

Event entry time prediction in financial business processes using machine learning: A use case from loan applications

Michael Frey	Andreas Emrich	Peter Fettke	Peter Loos
German Research Center for Artificial Intelligence and Saarland University	German Research Center for Artificial Intelligence and Saarland University	German Research Center for Artificial Intelligence and Saarland University	German Research Center for Artificial Intelligence and Saarland University
Michael.Frey@iwi.dfki.de	Andreas.Emrich@iwi.dfki.de	Peter.Fettke@iwi.dfki.de	Peter.Loos@iwi.dfki.de

Abstract

The recent financial crisis has forced politics to overthink regulatory structures and compliance mechanisms for the financial industry. Faced with these new challenges the financial industry in turn has to reevaluate their risk assessment mechanisms. While approaches to assess financial risks, have been widely addressed, the compliance of the underlying business processes is also crucial to ensure an end-to-end traceability of the given business events. This paper presents a novel approach to predict entry times and other key performance indicators of such events in a business process. A loan application process is used as a data example to evaluate the chosen feature modellings and algorithms.

1. Introduction

The financial crisis of 2008 has led to a dramatic turnover in the finance industry. Regulations such as Sarbanes-Oxley [1], Basel III [2], the Dodd-Frank Wall Street Reform and Consumer Protection Act [3] or the upcoming Basel IV severely transformed the way how risks are assessed and processes are organized in finance industry [4]. As many research concentrates on the risk assessment from a monetary standpoint, only few work addresses the relationship between compliance of business processes in finance industry and the associated risks.

For the compliance of business processes, the theory of Business Process Management (BPM) is key to an understanding of the temporality of business process events and their link to respective key performance indicators. Quite often, these are directly deduced or at least dependent on process performance indicators (PPI), a special kind of KPI that are modelled together with the business process [5] and are acquired throughout Business Activity Monitoring (BAM).

As financial risk is usually assessed by analyzing key performance indicators, the utilization of business

activity monitoring and the prediction of the associated PPI can be beneficial to get an early risk estimation.

The goal of this paper is to design a general approach for the prognosis of process performance indicators (PPIs) based on process mining technology [6, 7, 8]. In this paper, event entry times will be analyzed as a sample PPI in loan applications. The approach utilizes regression algorithms to predict real-value PPI. It is applied using several regression algorithms. The event log data from loan application interactions with customers are used as process instance data.

The research in this paper follows a design science paradigm, which aims at gaining scientific insight through the creation and evaluation of design artifacts [9]. Following the seven guidelines defined by [10], we develop a design artifact in form of a model and a prototypical implementation of a prediction system for event entry times. The knowledge base in design science consists of foundations (development of theories, constructs, models and frameworks) and methodologies (formalism, measures and validation criteria)[10]. As knowing the state of a future step can be beneficial in controlling a business process, our proposed approach of predicting a key performance indicator with machine learning is a novel solution to a relevant business problem.

The rest of this paper is organized as follows: In Section 2 we establish the state of the art by analyzing the research that has already been conducted in using key performance indicators for predictive analytics. In Section 3 we derive the requirements for our approach based on the previously established state of the art. Section 4 describes our developed approach and presents a use case that will show its feasibility and benefit. Again following the design science paradigm we will evaluate our artifact in Section 5 using data from a loan application process. In the final section we present our conclusion and give an outlook on future research.

2. Related Work

Existing work on compliance of processes in the finance industry did only scarcely consider entry times of business activity related events. In section 2.1 we describe existing key process performance indicators relevant for business process management, whereas section 2.2 addresses specific compliance approaches in the finance industry.

2.1. Key & process performance indicators

A KPI typically consists of a mathematical formula that converts observed values in a certain time period into a single real value, which is measured against a target value that in turn is linked to a specific business goal [11]. A decision maker or Business Process Engine can then use this information to control a running process or change the underlying process model to influence future processes. The majority of KPIs are of a financial nature with examples like *return of investment* or *sales volume*. However looking at recent research in BPM and BAM shows that KPIs also play an important role on the operational level [5, 12], where they can be divided in 3 categories, whether they concern time, cost or quality [13]

- How long did it take to finish a process/ process step?
- How much did the execution of a process/ process step cost?
- the failure rate of a process/ process step?

Most companies however struggle with the task of identifying their KPIs [14] let alone their underlying PPIs.

Schwegmann et al. [15] developed a tool named preCEP which adds predictive capabilities to an existing business activity monitoring platform, by facilitating complex event processing methods and key performance indicators. Their concept is built around an event processing network which transforms events, based on historical and current data, into meaningful KPIs. This is performed by a series of event processing agents (EPA), each executing a different operation like e.g. observing, filtering or pattern recognition, until they are transformed into numeric or categorical metrics which are then used to perform the actual prediction. The analysis component of preCEP utilizes a support vector machine for classification and regression and is built upon an existing design component based on the eclipse editor.

[16] propose a concept for predicting key performance

indicators in balanced scorecards, a performance measurement system that supplements traditional financial measures by considering different perspectives. The proposed concept utilizes association rules and a neural network in parallel to discover relations between different key performance indicators which are then used to predict future values.

2.2. Process compliance in finance industry

Business Process Compliance (BPC) follows the notion, that a business process instance follows its process model [17]. The process model itself often comprises organizational and domain knowledge, as well as regulatory requirements. In terms of finance industry, often risks are associated with a certain task. The approach of Sadiq et al. [16] associates risks with process activities in terms of so-called control objectives, which are monitored throughout process execution. The approach of Ghose & Koliadis [18] uses semantics to define the interlinking of those concepts and formulates rules on how to react to such KPI changes. The approach of Ly et al. [19] follows a similar direction and defines the structure in rule graphs. Becker et al. [20] combine process models and other resource models such as data models also via graph structures and apply pattern matching algorithms to determine impacts for risk assessments. Newer work by Lohmann [21] proposes a compliance by design approach that constructs process models and the other associated enterprise models in such a way, that no additional expertise is needed to explicitly specify the compliance rules, i.e. the compliance rule formulation is part of the design of the process. However, none of the depicted approaches directly addresses the timeliness of business process events and their impact on over-time assessments of the risk over time. Our evaluation of KPI prediction hence concentrates on this aspect.

3. Requirements for PPI prediction in business processes

As good as it is to know how a process is running at the current time, it is more beneficial to predict how a process will behave in the near or far future [22]. The majority of research concerning this process prediction focuses on predicting the result of a process or the next step in a process (fig. 1), where there are multiple possibilities available [23, 24, 25]. Evermann et al. [26] transferred the applicability of neural networks from Natural Language Processing to the domain of Next Step Prediction in Business Processes. This use case is especially useful for resource planning as a production resource may be used for a different process when

it is foreseeable that the current process will not be needing it. Process Performance Indicators are a good foundation for this kind of process prediction, as they deliver a reliable picture of the current process state.



Figure 1. Predicting the next step

In contrast to these approaches we do not just want to predict the next step but solve the following BPM problems, we derived by extending the aforementioned example questions:

- When will a future activity start?
- How long will a future activity take?
- How much will a future activity cost?

To answer these questions, instead of predicting what the next step is going to be or what the overall outcome of a process looks like, we assume that these questions are already answered and that we need to know how the KPIs for a future process step behave. Fig. 2 illustrates that we do not predict step c but the relevant PPIs of step c.

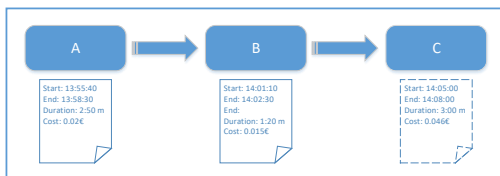


Figure 2. Predicting KPI

Given these problems and what we have learned from the above state of the art in predicting key performance indicators we derive the following requirements for our proposed approach.

Requirement 1 The approach should utilize historical event data and the related key performance indicators to create regression models for a generic scenario. From looking at similar approaches we can derive that it is necessary to train a prediction model on a certain amount of historical data from the target system.

Requirement 2 The approach should combine multiple regression algorithms to be able to utilize the one that best fits the current data. It also needs

to detect and select the essential features from the given dataset. Depending on the dataset different regression algorithms may have better performance.

Requirement 3 It is required that the approach can convert certain features into metric or binary values to be able to handle non metric features, e.g. categorical values. As the majority of the input features is of categorical nature, there has to be a functionality to convert them into a data format that is usable for the regression algorithm.

Requirement 4 The approach should be independent from any process modelling language. Given the different complexity of processes on different levels the utilized modeling language may vary. As the approach of Evermann et al. [26] shows, no process model at all may be necessary.

4. Proposed approach

Our approach revolves around the prediction of a key performance indicator of a process event, based upon information about the event and previous events. To showcase our approach we have a use case from the loan application field.

4.1. Use case

The use case we present in this paper is the filing and subsequent handling of loan applications in a financial institute and is provided by the BPI Challenge 2017 [27]. The process is roughly the following:

1. A customer uses a web interface to start applying for a loan, providing the goal of the loan and the desired amount. This action triggers an event that has the application system as its source. After providing the necessary information the customer submits the application which in turn triggers another event stating the fact that the application has been submitted.
2. Usually the next event originates from the internal workflow system of the financial institute and is related to the activity of handling the loan leads, i.e. the potential customer.
3. After the application is accepted, which is indicated by a specific event (A_Accepted), one or multiple offers are created. Offer related events originate from another source, called *offer*. (The ability to create multiple offers for one application was implemented because of feedback from the

BPI challenge 2012 which featured data from the same financial institute)

4. From here on out the process differs a lot from application to application. In some cases there is a lot of back and forth communication about offers or missing documents which triggers a series of event with the lifecycle transition *suspend* and *resume* for the currently running activity. During this time the application is still being validated in the workflow system.
5. Finally either one offer is accepted, all offers are rejected or the whole process is just canceled

4.2. Data source

The data [28] was provided in the *Extensible Event Stream (.xes)* format, a XML-based format for the interchange of event logs. It consists of 1,202,267 events that belong to all 31,509 loan applications that were filed in the year 2016 and handled up until February 2nd 2017. The events have the following list of attributes, from which we chose the features in our regression algorithms:

application_id A unique identifier for the loan application (current trace) the event belongs to.

event_id The unique identifier for the event itself.

name A standard attribute of the Extensible Event Stream file format that is used to describe the name of the related activity. In this case, the name starts with a capital letter denoting the system in which the event originated. Separated by an underscore follows the name of the related activity, e.g. *A_Create Application*. The value is therefore of categorical nature.

action This categorical attribute denotes if the related activity was originally created, deleted, obtained by a new user or released by a user, so another one can obtain it.

user This attribute holds the unique identifier of the user who triggered the event.

transition Another standard categorical attribute of the XES format, which denotes the lifecycle transition of the related activity. The possible values are *complete*, *start*, *suspend*, *schedule*, *withdraw*, *ate_abort* and *resume*. The transition *withdraw* explains that the resource assignment of the activity

has been revoked and *ate_abort* stands for the cancellation of the execution.

origin A categorical value that denotes the system from which the event originates. There are three different systems involved in the loan application process: *Application*, *Workflow* and *Offer*. The majority of events originate from the Workflow system (63.94%).

timestamp The time when the event occurred. In this dataset, the Unix timestamp is measured in elapsed milliseconds since January 1 1970.

application_type A categorical value that denotes whether the application is for a new loan (89.24%) or to modify the conditions of an existing loan (10.76%), e.g. to raise the current amount or term.

loan_goal A categorical value that denotes the goal the customer wants to achieve with the loan. The possible values are a set of predefined labels, chosen from typical cases. With 9328 cases (29.60%) cars are the most common reason for a loan application, followed by home improvement (24.33%) and existing loan takovers (17.77%).

requested_amount An integer value that represents the sum that the customer requests for the loan. The average amount for loans was 16233€.

4.3. Overview

As mentioned before we want to predict the timestamp of the event that will occur next or rather the elapsed time since the first event from a trace until the next event occurs.

$$t_{elapsed}(x) = t_x - t_0 \quad (1)$$

Formula (1) shows how we calculate the elapsed time of the next step (x) based on the first timestamp of the trace (t_0).

The reason we consider this value and not just the difference between two events is because of the linearity of the value. As we want to use linear regression to predict the values we would run into issues with the coefficients when predicting the duration of the next step based on the duration of the current step. All following steps were implemented in *python* using the well-known *pandas* framework for data handling and *sklearn* for preprocessing, training and evaluation.

4.4. Data preprocessing

Regression takes a vector of values as input to predict a single value, so one of the first steps of data

preprocessing is to prepare the given data in a way that creates such a feature vector. Additionally we have to create the labels, i.e. the values we want the regression to predict. We executed the following steps to prepare our data for linear regression.

1. The first step of preprocessing is converting the timestamps so they represent the elapsed time relative to the first event of the respective trace. We do this by using the above mentioned equation (1) for every row in the dataset. After this step every row in our dataset has a value that depicts the elapsed time since the first event of the trace in milliseconds.
2. Next we add the time that elapsed until the next event by iterating over our dataset and adding the newly created *elapsed_time* feature of the next row ($x + 1$) to the current row (x). This feature will later become the labels, we want to predict. As they do not have a next event, we remove the last event for every trace.

event	elapsed_time	elapsed_time_next
x	89647125	91908995
x+1	91908995	92268690

Figure 3. Elapsed time of next step in seconds

3. As regression only works with features that are either metrical or at least dichotomous we cannot use a number of our features directly as they are of categorical nature. However simply encoding them into numbers is not sufficient as they would still not be dichotomous. To still be able to use these features we utilize a technique called *dummy coding*. Dummy coding creates multiple binary variables to represent the categorical values of a feature. Given k different values in a feature, dummy coding will create $k-1$ new features. Figure 1 illustrates dummy coding for the origin feature. As we do have multiple categorical features with up to 26 different values this results in a total of 77 features we use for the regression.

Origin		Origin_Application	Origin_Workflow
Application	→	1	0
Workflow		0	1
Offer		0	0

Figure 4. Dummy coding of origin feature

4. The last step of preprocessing is to prepare the data for cross validation. Cross validation is a technique where you split the dataset into a training set, which is used to train the regression model and a test set that can later be used to evaluate the resulting model. This is done to minimize overfitting, where the model only performs well on this specific dataset, but not on independent datasets. As we do not want to separate events that belong to the same application we randomly split the list of applications and use the resulting sets to select all events that belong to either one. We selected 80% of the values for the training set and the rest for the test set.

This completes the data preprocessing and the resulting datasets can be used for regression.

4.5. Regression algorithms

To receive a more robust result from our analysis we decided to use multiple regression algorithms to find the one that best fits our dataset. We chose the following four regression algorithms.

4.5.1. Multiple linear regression. Linear regression gets its name from the fact that the dependent variable y is a linear combination of the regression coefficients β (2). If there are multiple independent variables x it is called multiple linear regression. During the training of the regression model, these coefficients are calculated and recalculated until the model best fits the training data. Hence the coefficients are the variables of the regression model.

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i \quad (2)$$

Using linear regression with our feature set results in the following underlying equation (3). We did not use the attribute user as it would create too many dummy variables and thereby too many features for some of the algorithms. After the preprocessing we are left with the following variables (Table 2) we can use with the regression algorithm, where x is used to depict a variable that references the current event and $x + 1$ references the next event, e.g. $\text{name}(x + 1)$ is the name of the next event in the sequence. The variables for name, action, transition, origin, application_type and loan_goal resulted from the dummy coding and are therefore binary values representing the previous categorical values:

Feature	Variables
current event	x
next event	$x + 1$
name(x)	$a_1, \dots, a_m; m = 25$
action(x)	$b_1, \dots, b_n; n = 4$
transition(x)	$c_1, \dots, c_o; o = 6$
origin(x)	$d_1, \dots, d_p; p = 2$
application_type	e
loan_goal	$j_1, \dots, j_r; r = 13$
name(x+1)	$k_1, \dots, k_s; s = 24$
requested_amount	u
elapsed_time(x)	l
elapsed_time(x+1)	y

Table 1. Variable definition

$$\begin{aligned}
y_1 = & \beta_0 + \beta_1 a_{1i} + \dots + \beta_{25} a_{mi} \\
& + \beta_{26} b_{1i} + \dots + \beta_{29} b_{ni} \\
& + \beta_{30} c_{1i} + \dots + \beta_{35} c_{oi} \\
& + \beta_{36} d_{1i} + \beta_{37} d_{pi} + \beta_{38} e_i \\
& + \beta_{39} j_{1i} + \dots + \beta_{51} j_{ri} \\
& + \beta_{52} u_i + \beta_{53} l_i \\
& + \beta_{54} k_{1i} + \dots + \beta_{77} k_{si} + \epsilon_i
\end{aligned} \quad (3)$$

Simple linear regression utilizes the *ordinary-least-squares* method to estimate the unknown parameters. This method calculates the *error sum of squares (ESS)* for every candidate b for every regression coefficient β , choosing the one with the smallest sum as the best fit. The formula given in (4) describes how the error sum is calculated by measuring the distance between each data point (x_i, y_i) and the hyperplane ($y = x^T b$):

$$S(b) = \sum_{i=1}^n (y_i - x_i^T b)^2 \quad (4)$$

4.5.2. Ridge regression. If ordinary least squares cannot find a regression coefficient or multiple values for one coefficient the regression model is either over-fitted or under-fitted. Ridge regression utilizes 2 regularization which adds a bias towards smaller values to select the coefficients.

4.5.3. Lasso. Lasso or *least absolute shrinkage and selection operator* is another regression method that utilizes regularization to improve the selection of coefficients. Additionally it performs feature selection to improve the regression model. A problem with lasso

is, that it has limitations when it comes to data with high dimensionality and not enough examples, where it only choses at most as many variables as there are examples.

4.5.4. Elastic Net. Elastic net is a regression method that utilizes a combination of the L1 and L2 regularization from lasso and ridge regression. It overcomes the dimensionality limitations of lasso by introducing an additional quadratic penalty.

To prevent single sequences of events from tainting our results we also create extra regression models for every pair of event type and next event type we have in our data. Doing this allows us to identify the pairs where the regression algorithms struggles and also gives us models that better fit individual pairs.

5. Evaluation

In this section we perform the evaluation of our approach based on the selected dataset from the BPI Challenge 2017 utilizing widely accepted metrics.

5.1. Evaluation data

As mentioned earlier we utilized cross-validation with the training and test datasets we created earlier to improve our evaluation and strengthen the robustness of our approach.

5.2. Metrics

5.2.1. MAE. The *mean absolute error* is a measure of difference between the real measured values and the predicted values. As the name suggests it is the average of how far the predictions differ from the observed values. The MAE may not be as good a metric for prediction as e.g. the RMSE but it is easy to interpret.

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n} \quad (5)$$

5.2.2. RMSE. Although the RMSE is often interpreted in the same way as the MAE it should not be. Because the RMSE is the square root of the squared errors, large errors have a disproportional influence of its value. This makes the RMSE very sensitive to outliers.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - x_i)^2}{n}} \quad (6)$$

5.2.3. R squared. The *coefficient of determination* is a number that indicates proportion of the variance in the independent variable. In linear regression it is the

square of the sample correlation coefficient between the observed and predicted values.

5.3. Results

Figure 5 through 8 showcase the values predicted by the regression model on the vertical axis and the actual value on the horizontal axis respectively. As the values are measured in milliseconds the scale reaches from 0 to 1.2×10^{10} (roughly half a year) on the horizontal axis and around the same value on the vertical axis. The major difference between the horizontal and the vertical axis is, that the vertical axis also has negative values. Those result from prediction errors where the result can be negative values. We decided to scale the axis including the negative values to better highlight the differences between the four algorithms.

	Linear	Ridge	Lasso	Elastic net
MAE	63961s	63882s	63961s	130405s
RMSE	200009s	196177s	200009s	329331s
R^2	0.9596	0.9612	0.9596	0.8907

Table 2. Evaluation metrics

Table 2 displays that the ridge regression algorithm has the best results for all three evaluation metrics and is thereby the best fit for our dataset. There is nearly no difference between linear regression and Lasso regression. This could be because of the relatively high amount of features in the dataset. ElasticNet has the worst results, which can be traced back to the fact that it did not converge because of the high amount of variables.

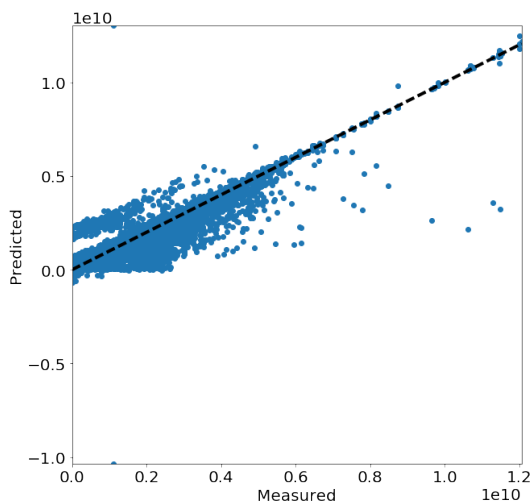


Figure 5. Linear regression

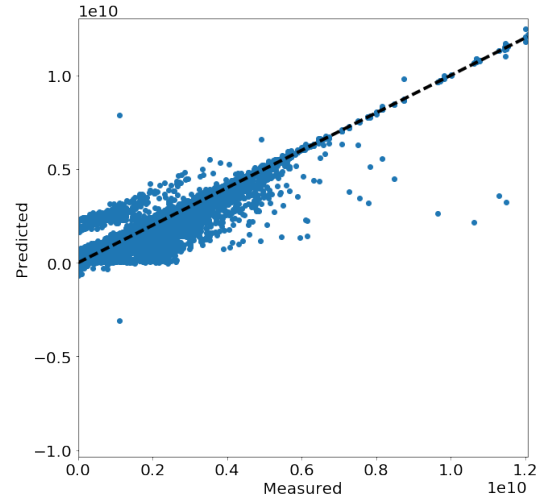


Figure 6. Ridge regression

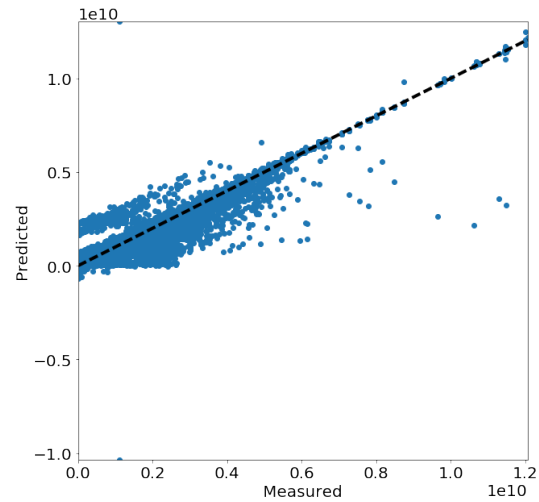


Figure 7. Lasso regression

The scatter plots clearly show that ridge regression is more robust against outliers and that ElasticNets tends to predict mostly values that are too low. It is also noticeable that the prediction accuracy rises with higher values.

To sum up our results, ridge regression is the best algorithm to predict the elapsed time between the first event and the next event. Although our prediction is off by around 17 hours, it is still a good result considering that the mean of all standard deviations of the target variable y is around 37 hours and the R^2 score for ridge regression shows that the variance can be explained through the regression model.

The above results show that it is possible to predict the time when the next event occurs given the knowledge of what the next event would be.

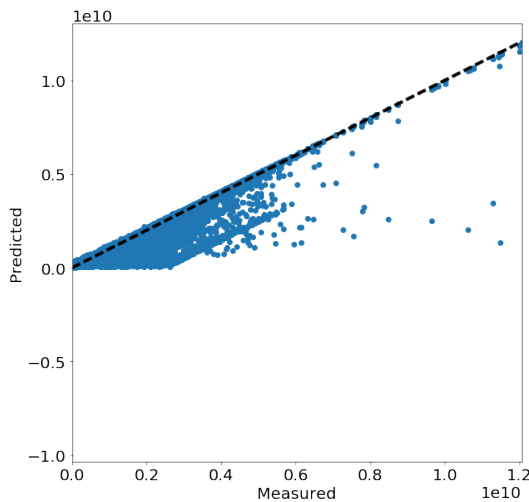


Figure 8. ElasticNet

5.4. Discussion

The data clearly indicates, that the accuracy of the regression is in general very high. The different regression models do not expose significant differences regarding the metrics, however ridge regression slightly outperforms the other models, whereas ElasticNets show less accurate results than the other regression models.

Overall, the variance in the data set is quite high. Some loan application processes have been halted for almost a month, causing significant delays in the underlying process instance and, from a data perspective, constituting an outlier that heavily impacts the model learning. The data itself contain no evidence, why the delays occurred, but most likely external events like a customer not responding or a back office worker calling in sick could lead to such delays. In practice, such external shocks cannot be entirely modelled in the respective regressions, but must be handled in the risk & compliance management processes of the given company.

A weakness of this approach follows from the relatively high amount of feature variables that result from the necessary usage of dummy coding when working with data that largely consists of categorical values. The more feature variables are used, the more overfitting will happen in the trained model.

The data prove that events in a financial business process can be predicted with a high accuracy. This will provide new opportunities to assess financial risks in a workflow-oriented view and to gain actionable insights on single transactions such as loan application cases. In a similar manner, arbitrary key performance

indicators or process performance indicators could be predicted, which could be directly linked to existing risk assessment frameworks. The main challenge is to identify such performance indicators in the first place.

6. Conclusion and outlook

This paper has shown a novel approach for predicting start and end events of business process activities. The use case has been evaluated in a finance industry use case, a loan application process.

The approach itself proves feasible and robust with high accuracy. However, the given use case example clearly shows, that many outliers in the data exist and hence limit the single case accuracy severely. As this is rooted in many different external and process-intrinsic reasons, e.g. as an extended waiting for a customer response or vacations of the respective back office workers, additional data should be included to make the regression models even more accurate. This additional data could be acquired by implementing techniques for semantic preprocessing [29] or by adding additional data sources like an availability calendar for employees. Furthermore as outliers represent cases of potential special interest, further research could be done to investigate if the data holds information about the causality of such cases. One could also treat outliers as completely different cases and train additional models for these cases.

For finance industry, the approach shown in this paper could help to interpolate financial risks on a timeline and hence improve the risk value assessments for a given point in the future. This could be very valuable for loan processes or credits of any other kind. Moreover, the approach depicted in this paper can be transferred to business processes in general, providing new means for providing traceability and checking compliance of a business process throughout its execution.

Acknowledgement: The research described in this paper was partly supported by a grant from the German Federal Ministry of Education and Research (BMBF), project name: REPAIR, support code 01|S12050. Moreover, the authors would like to thank the anonymous reviewers and the mini-track co-chairs for their constructive comments which contributed to the improvement of this article.

References

- [1] J. D. Piotroski and S. Srinivasan, "Regulation and bonding: The sarbanes-oxley act and the flow of international listings," *Journal of Accounting Research*, vol. 46, no. 2, pp. 383–425, 2008.

- [2] "Basel 3 Bank for International Settlements." Available at <http://www.bis.org/publ/bcbs189.htm>.
- [3] U. H. of Representatives, "Dodd-frank wall street reform and consumer protection act. public law 111-203 [hr 4173]," 2010.
- [4] N. Dick, *Choosing innovation over cutting costs: Market opportunities and challenges in the insurance industry*. Verlag Versicherungswirtschaft., 2007.
- [5] A. Del-Río-Ortega, M. Resinas, and A. Ruiz-Cortés, "Defining Process Performance Indicators: An Ontological Approach," *On the Move to Meaningful Internet Systems: OTM 2010*, vol. 6426, no. Confederated International Conferences: CoopIS, IS, DOA and ODBASE, pp. 555–572, 2010.
- [6] W. M. P. van der Aalst, H. A. Reijers, A. J. M. M. Weijters, B. F. van Dongen, A. K. Alves de Medeiros, M. Song, and H. M. W. Verbeek, "Business process mining: An industrial application," *Information Systems*, vol. 32, no. 5, pp. 713–732, 2007.
- [7] D. Werth, A. Emrich, and P. Loos, "Zentralisierte Echtzeit-Steuerung von Geschäftsprozessen in der Logistik," in *Wandlungsfähige Produktionssysteme. HAB-Forschungsseminar, October 8-9, Hannover, Germany*, 2010.
- [8] C. Houy, P. Fettke, and P. Loos, "Empirical research in business process management analysis of an emerging field of research," in *Business Process Management Journal*, vol. 16, pp. 619–661, 2010.
- [9] T. Wilde and T. Hess, "Methodenspektrum der Wirtschaftsinformatik: Überblick und Portfoliobildung," *Arbeitsbericht 2006/2: Institut für Wirtschaftsinformatik und Neue Medien der Ludwig-Maximilians-Universität München*, no. 2, pp. 1–14, 2006.
- [10] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design Science in Information Systems Research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.
- [11] H. J. Schmelzer and W. Sesselmann, *Geschäftsprozessmanagement in der Praxis*, vol. 91, 2008.
- [12] B. Wetzstein, Z. Ma, and F. Leymann, "Towards measuring key performance indicators of semantic business processes," *Business Information Systems*, pp. 227–238, 2008.
- [13] M. Z. Muehlen and M. Rosemann, "Workflow-based process monitoring and controlling-technical and organizational issues," *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, vol. 00, no. c, pp. 1–10, 2000.
- [14] M. Badawy, A. A. El-Aziz, A. M. Idress, H. Hefny, and S. Hossam, "A survey on exploring key performance indicators," *Future Computing and Informatics Journal*, vol. 1, no. 1-2, pp. 47–52, 2016.
- [15] B. Schwegmann, M. Matzner, and C. Janiesch, "preCEP: Facilitating predictive event-driven process analytics," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7939 LNCS, pp. 448–455, 2013.
- [16] A. M. A. El-mongy, A. E.-d. Hamouda, N. Nounou, and A. W. Abdel-moneim, "Design of Prediction System for Key Performance Indicators in Balanced Scorecard," *International Journal of Computer Applications*, vol. 72, no. 8, 2013.
- [17] S. Sadiq, G. Governatori, and K. Namiri, "Modeling control objectives for business process compliance," *Business Process Management*, pp. 149–164, 2007.
- [18] A. Ghose and G. Koliadis, "Auditing business process compliance," *Service-Oriented ComputingICSOC 2007*, pp. 169–180, 2010.
- [19] L. T. Ly, S. Rinderle-Ma, D. Knuplesch, and P. Dadam, "Monitoring business process compliance using compliance rule graphs," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7044 LNCS, pp. 82–99, 2011.
- [20] J. Becker, P. Bergener, P. Delfmann, and B. Weiß, "Modeling and Checking Business Process Compliance Rules in the Financial Sector," in *ICIS*, p. 12, 2011.
- [21] N. Lohmann, "Compliance by design for artifact-centric business processes," *Information Systems*, vol. 38, no. 4, pp. 606–618, 2013.
- [22] C. Houy, P. Fettke, P. Loos, W. M. P. van der Aalst, and J. Krogstie, "BPM-in-the-Large Towards a Higher Level of Abstraction in Business Process Management," *E-Government, E-Services and Global Processes*, vol. 334, no. October 2015, pp. 233–244, 2010.
- [23] M. Polato, A. Sperduti, A. Burattin, and M. de Leoni, "Time and activity sequence prediction of business process instances," *arXiv preprint arXiv:1602.07566*, 2016.
- [24] W. M. P. Van Der Aalst, M. H. Schonenberg, and M. Song, "Time prediction based on process mining," *Information Systems*, vol. 36, no. 2, pp. 450–475, 2011.
- [25] A. Rogge-Solti and M. Weske, "Prediction of remaining service execution time using stochastic petri nets with arbitrary firing delays," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8274 LNCS, pp. 389–403, 2013.
- [26] J. Evermann, J.-R. Rehse, and P. Fettke, "Predicting process behaviour using deep learning," *Decision Support Systems*, vol. 100, pp. 129–140, 2017.
- [27] "BPI Challenge 2017." Available at <https://www.win.tue.nl/bpi/doku.php?id=2017:challenge>.
- [28] "Data BPI Challenge 2017." Available at <http://data.4tu.nl/repository/uuid:5f3067df-f10b-45da-b98b-86ae4c7a310b>.
- [29] A. Emrich, D. Werth, and P. Loos, "Real-Time Semantic Process Change Impact Analysis," in *Innovative Unternehmensanwendungen mit In-Memory Data Management 2011. Innovative Unternehmensanwendungen mit In-Memory Data Management (IMDM-11), December 2, Mainz, Germany*, pp. 91–100, 2011.